DOCUMENT RESUME

ED 064 398 TM 001 646

AUTHOR Roid, G. H.
TITLE Systems Design for Course Evaluation.
PUB DATE Apr 72
NOTE 43p.; Paper presented at the AERA symposium, "Research, systems design and the future of university course evaluation", (Chicago, Ill., April 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Course Evaluation; Course Objectives; *Evaluation Techniques; *Measurement Instruments; *Models; Questionnaires; *Systems Approach; Test Results; Tests

ABSTRACT

It is the thesis of this paper that much of the present concern over course evaluation has come about because evaluation has become synonymous with the use of a single questionnaire instead of the broader process of evaluating the course as an educational program. An attempt is made to redefine course evaluation in light of recent work on educational evaluation, and suggest a model for a course evaluation system. First, terms are defined and the aspects of a course which should be evaluated are delineated. This "ideal" is compared to existing course evaluations and the need for new emphases are explained. A systems design is then presented. Evaluation is defined "as a process of examining certain objects and events in the light of specified value standards for the purpose of making adaptive decisions." Current course evaluation questionnaires do not specifically relate to specific goals and standards for a particular course. A model evaluation system begins with a specific definition of the purpose of the course evaluation by all those who will use the information and judgments made public from it. An assessment of whether real and significant changes will be possible if evaluation is conducted must be guessed. Next, the various subgroups served by the course are identified. The intended goals of the course are specified and an initial list of intended inputs, processes, and outputs for a course is drawn up. Observation methods, tests, checklists, simple frequency counts, and other measures which can be pre-designed are constructed. Findings, data, and evaluative judgments are written up in report form and a mini-experiment can be run by taking evaluation reports to decision-makers and discussing results with them. (Author/CK)

ED 064398

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEH REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

4,5

Systems Design for Course Evaluation

G. H. Roid

Centre for Learning and Development

McGill University

TM 001 646

Paper presented as part of the symposium, "Research, systems design and the future of university course evaluation", at the meeting of the American Educational Research Association, Chicago, April, 1972.

Systems Design for Course Evaluation

G. H. Roid
Centre for Learning and Development
McGill University

Most universities and colleges in North America employ course and instructor rating systems. These typically involve a questionnaire given to students at the end of courses, the results of which are tabulated, summarized and sometimes published or distributed by student groups or test bureaus.

Despite the current widespread use of course evaluation in colleges and universities, concerns about the validity and usefulness of questionnaires and rating systems still remain.

In terms of validity, it is not clear that currently popular rating scales or questionnaires are valid measures of the things they are intended and expected to be. Are they adequate measures of teaching ability or course effectiveness? Do they give students valid information to choose courses? Are they valid measures for use in the promotion of faculty members? Do they tell us how much students change, grow or learn from a course? It seems that we presently do not know the answers to these questions, and many suspect that the answers are "not very well". Certainly we can always improve in the evaluation of such complex entities as university courses.

In terms of usefulness, the author, who works with users of ratings (e.g., an instructor trying to improve his course or a dean evaluating his department's teaching) has found that questionnaire results are not always easily translated into meaningful course improvement or behavior change in teachers. Some users do not understand computer-printed results, others find questionnaire results and the questions themselves vague or irrelevant to their particular courses. General ratings of satisfaction or dissatisfaction do not pinpoint specific aspects of a course which need changing. In many cases changes in basic attributes of a course were not possible to begin with and questionnaire results become a thorn in the side. Sometimes results come too late to benefit decision making. Given the fact that rating scales have been used since 1926, and widespread student course surveys for nearly a decade, it is surprising that there is little documentation of the usefulness of them. To the author's knowledge little seems to be known about how instructors, students or administrators actually use results from surveys. If we take seriously the argument of Cronbach (1963), that evaluation should be keyed to course improvement we must measure our investment in course evaluation in terms of actual improvements that have come about.

It is the thesis of this paper that much of the present concern over course evaluation has come about because evaluation

has become synonymous with the use of a single questionnaire instead of the broader process of evaluating the course as an educational program. The use of questionnaires alone has constrained us to examine only a narrow range of course attributes and outcomes. Much of past research has focused on the psychometric properties (reliability, validity, norms, factor structure, etc.) of evaluation questionnaires. Recent developments in evaluation theory would suggest that the questionnaire is only part of an evaluation process or system which includes initial specification and evaluation of objectives, measurement of outcomes and mechanisms for using evaluation data as corrective feedback. This paper attempts to redefine course evaluation in light of recent work on educational evaluation, and suggest a model for a course evaluation system. First, terms are defined and the aspects of a course which should be evaluated are delineated. This "ideal" is compared to existing course evaluations and the need for new emphases are explained. A systems design is then presented.

Defining Course Evaluation.

Evaluating an educational program can be a very informal or a very complex process. An instructor can decide to change a course on the basis of a talk with a single student or he could launch an involved study to measure the detailed effectiveness

of the course on a large number of students before deciding. Evaluation involves both making observations and rendering judgements. Paulson (1970) has provided a more complete definition:

> "Evaluation is defined as a process of examining certain objects and events in the light of specified value standards for the purpose of making adaptive decisions. The crucial dimension of this definition is the assigned task of providing relevant and valued information which may serve the decision process. Our concern is to provide information to a decision-making body specifically related to a given value and that will subsequently improve the quality of decisions to be made. This does not imply that evaluation will insure perfect decisions, but rather that decisions based upon appropriate data will be improved."

Applying this definition to course evaluation is revealing. Course evaluation would be seen as the process of examining the people, materials and events associated with a course in light of what goals, standards and expectations are set for it for purposes of making decisions which improve the course.

Perhaps the most important attribute of this definition is that it describes evaluation as criterion-referenced rather

than norm-referenced. Observations in courses are related to standards not observations in other courses. The accomplishments of a course (e.g., student learning) are related to what the course was intended to achieve. Instructors are not compared with one-another, but with criteria and standards (which they help to define).

Measuring Traits vs. Descrepancies.

In contrast to what is implied by the definition, current course evaluation questionnaires do not specifically relate to specific goals and standards for a particular course. Rather they are usually a set of *a priori* statements about what constitutes good teaching and practices (e.g., "The instructor encouraged the students to express opinions"). As Stake (1967) has said of educational evaluation in general, "Little attempt has been made to measure the descrepancy between what an educator intended to do and what he did."

Course evaluation questionnaires are oriented toward the measurement of traits of instructors, course methods and materials rather than assessing the accomplishment of goals in specific terms. Students are sometimes asked to rate the instructors "friendliness", "rapport", or "teaching ability". They are asked about the "relevance" of the course or how "interesting" the text was. These ratings may be of some value, but they

are neither specific nor always related to what the course is expected to accomplish. For example studies of the relationship between student ratings and student examination results (a measure of accomplishment?) have revealed only moderate (.30-.40) correlations (e.g., Cohen and Berger, 1970).

The vaguely stated traits of one instructor or course are compared with another or with a norm for a department or university. This procedure is roughly analogous to grading on a curve.

In contrast, the alternative is to define for each course what its specific goals are in terms of behavior the student is expected to be able to do before, during and after the course. Also, it would be helpful to define what the instructor, his staff and his learning materials will be doing and accomplishing during the course. Even if these objectives were only described after the course by examining what happened that was not planned, it would provide a more relevant basis for evaluation. Ratings, observations and other measures of the degree to which these objectives were reached would constitute the basic descriptive data for the evaluation.

Importance of Decision Makers.

Another important aspect of the definition of evaluation presented above was its emphasis on providing useful information

for people who make decision about the course.

A common example might be helpful here. Have you ever had bad service from a large government agency? Have you ever felt that arguing about improving that service with a clerk in the agency might accomplish little? If you have you have probably realized that you would have to talk to the people who really make decisions in that agency. Also, you would have to be convincing. Your discussion would have to be information that the decision maker might actually use (e.g., a letter in addition to a talk) to improve his service.

Somewhat analogously, information for course improvement must be directly useful and convincing to decision makers who can change a course. The goal of providing information for course improvement is not to bring about perfect decisions, but to bring about decisions which are better than chance or better than those made without systematic information. To make this happen, the evaluation effort must involve consideration of the following questions:

1. Who makes decisions about improving the course? (The instructor? Departmental committees?)
2. What kinds of information and results would the decision maker(s) really use?
3. When do they need it?
4. If changes are not likely to be made, will the cost

of collecting information (time and money) be worth it? What are the costs of not collecting information?

5. What would be the consequences of negative evaluation or of significant changes?

Only when we can answer some of the questions can we feel motivated to spend lots of time evaluating systematically. After all, we all evaluate (make judgements) and make changes in courses most of the time anyway - on the basis of impressions from students, new research in the subject matter, or other information. What we need now is more complete information that is both useful and cost/effective in terms of staff and student effort as well as other costs. And no one can tell the decision makers exactly what that will be in every case. The decision makers have to tell the evaluators (even if they are the same people).

Aspects to Evaluate in a Course

Models of educational evaluation such as that of Stake (1967) or Stufflebeam (1968) are directly related to (and contributed to) the definition of course evaluation presented here. These models suggest that evaluation should be the describing and judging of an educational program (e.g., a course) in terms of the inputs, process, outputs and goals of that program.

The inputs to a course are the students that enter it,

and the materials, staff and planning which are invested in it. Courses are usually not meant for all students, but for a definable set with certain preparation, interests and needs. Basic resources like classrooms, lab equipment, books, media, library services, and a staff which includes appropriate teaching assistants or clerical helps are all inputs. All of these inputs should be determined to some extent by the goals of the course. They are going to aid or inhibit the teaching effectiveness of the course and therefore should be objects of evaluation.

The process in a course include all of the day to day interactions between the student and instructors, between students and students (e.g., in discussions or peer-teaching), and between the student and books, media or self-instructional materials. Process includes what the instructor(s) and students actually do during each phase of the course. The goals of a course tell us what kind of teaching methods (seminar, field-work, etc.) and processes we should set in motion. The nature and quality of these processes will perhaps have the greatest impact on course effectiveness.

The outputs from a course include the changes in students (new knowledge, skills, attitudes, etc.) which are consequences of the course. They may be short or long term, thought of in terms of one student or a whole group, related to other courses

in the university or to definable jobs or skills needed after graduation. Some outcomes are more difficult to define or more difficult to gather information on. This does not mean that they are indefinable but that it will take us longer to adequately define them. One way to achieve definitions of outcomes is to precisely observe and analyze the attributes of two kinds of people, (a) those that we feel have already achieved the outcomes we desire, and (b) those who have not. An observation technique which was constructed to distinguish between these two kinds of people would play a central role in output evaluation.

If we have limited time and resources to evaluate, it would seem wise to concentrate on outputs. We might adequately describe inputs and process but if we do not know what students gained from them we can not really know whether our investment in the course was justified. The importance of evaluating outputs is depicted in the following fable ; from Saslow (1970):

> Three magicians , accompanied by a simpleton find a pile of bones. The first magician successfully commands the bones to form a skeleton. The second covers it with flesh. The third announces that he can bring it to life. The simpleton tries to stop him, pointing out that it is a tiger. This does not stop the magicians, the simpleton climbs a tree and the magicians are killed.

> The moral could be phrased: "If you have not defined the output, how are you going to control the inputs and the costs?"

Once we have described what went into a course, what process actually went on during it, and what outcomes came out of it, we compare it to a standard. That standard is defined by the intents, goals and expectations of us and everyone who has a vested interest in the course. We ask ourselves whether the inputs and processes were appropriate for, and facilitated the outcomes.

The goals of a course are its most important reference points. They are statements of intended outputs. The literature on defining measurable educational goals and objectives is volumonous. The reader is refered to Yelon and Scott (1970) or Popham and Baker (1970). Briefly, objectives define what a person (the learner, the instructor, etc.) is intended to do, under what conditions, and to what degree. But the specification of goals does not simply mean putting whatever is currently done in a course into operational, measurable terms with conditions and criteria for completion. It is also important that an in-depth search be conducted for the ultimate goals and standards of the entire "audience" (students and society) served by a course. This involves the analysis of what needs exist in society at large and what tasks people

actually have to do to fill those needs. Geis (1970) gives an excellent discussion of this analysis:

> "In a world which had no tradition of education and educational institutions, an instructional designer might begin by examining the environment for needs or problems. Some of these might be alleviated through development and use of instruction.
>
> Presumably all teaching is aimed at providing people with the ability to do things so that they can affect their environment in certain ways. The output of instruction is not merely the skills and knowledge that people acquire but a supply of people who act upon the world to produce certain effects (including effects upon themselves).
>
> Education is a system to produce changes in the environment by developing mediators of such changes."

Geis goes on to give several examples such as-- the ultimate goal of a nursing training course might be to produce comfortable and healthy patients (served by nurses), or the goal of training language conversation skills is to allow the speaker to have specific effects on a listener.

This is not meant to imply that only a course designer defines goals. Goals come from the common ground discovered

between what students, society, or other educational programs require and what the instructor (or the "state of the art" in the subject matter) is able to provide. Everyone has a stake in defining the goals for university courses, particularly in these days of budgeting and growing limitations on resources.

Implied Deficiencies of Current Course Evaluation

The discussion up to this point has described an "ideal" to strive for in thinking about course evaluation. It reveals our relative lack of sophistication and knowledge at present with respect to a) contrasting intents and accomplishments rather than measuring traits, b) the problems of insuring that evaluation information is valued and used by decision makers to bring about real changes in instruction, c) the evaluation of the goals of the course themselves, and d) the measurement of learning outcomes as part of evaluation. These points have been discussed above to some extent, but the last two points deserve more elaboration.

Evaluating Goals. Making judgements about the current goals of a course should involve decisions about a) their clarity, b) the adequacy of their performance criteria, c) the appropriateness of the conditions specified, d) their meaningfulness, and e) the degree to which students were actually required to work towards them. Judgements about

clarity would be made by applying one of the checklists of attributes for well-defined objectives (e.g., Nelson and Paulson, 1969). Performance criteria are the guidelines for what constitutes an acceptable level of mastery or attainment of the goal. If these criteria are inappropriate by being too lax or too stringent or focus on trivial aspects of student behavior they will interfere with instructional effectiveness. For example, a research methods course for biologists might intend for students to be able to write research reports in a form understood by other biologists. If the criteria for "a well written report" only specified the organizational properties ("Problem", "Results", "Discussion", etc.) it would be incomplete. Students might not attend to other aspects of report writing. Appropriate criteria are probably only developed over time by successive approximation.

The specification of the conditions under which a goal is to be achieved may prove inadequate. For example, if a goal in a sensitivity training course focused on students' increased "statements of frank opinions" within the training group only, the ultimate goal of frank communication in other settings may not be facilitated. In other words one must ask himself "is this the right (or only) condition under which this behavior should occur".

Perhaps the meaningfulness of goals is the one thing we

need to evaluate more than any other. As Saslow (1970) has said, the evaluator "...should deal with a multiplicity of... sources in producing and rank-ordering objectives". "It is better to begin with as many alternatives as one can manage, rather than defining one's job as the 'behavioralizing' of whatever is presently being done, that is, assuming that traditional or status quo goals are valid." Anderson (1969) has taken a similar view (which also relates to the discussion below concerning learning outcomes):

> "Some people argue for empirical validation of instructional materials seem to take the position that effectiveness in modifying student behavior is the sole criterion for judging instruction. Let me emphasize that this is not my position. Lessons, units, and curricula should be judged in terms of the extent to which they reach their goals, but this cannot be the only criterion. Other criteria include the cost of the instructional sequence to students and teachers and any side effects (Stake, 1967). The accuracy, up-to-dateness, and elegance of the subject matter has been the important criterion for the prominent curriculum reform projects. A most important criterion is the worthiness of the goals the instruction aims to reach. As Scriven (1967) has noted, 'it is obvious that if the goals aren't worth achieving then it is uninteresting how well they were achieved'. A complementary assertion is also true: No matter how worthy the goals, a lesson cannot be valued highly if it is ineffective in reaching these goals. Effectiveness should be neither overrated nor underrated as a criterion for judging instruction."

Geis (1970) provides what are, perhaps, the most complete guidelines available for insuring that instructional goals are related to real needs for skills and knowledge in the environment. His analysis involves analyzing needs for changes in the

environment, defining the human role in these and analyzing the tasks that make up the human "job". Much needs to be done in this area. As he says, "In most academic areas not even traditional job analyses have been performed...It would not be impossible to determine what historians do; but such a specification of their activities does not exist... The activities of a good citizen, or an informed, thoughtful person remain even more mysterious."

And, certainly, all this effort is lost if during the actually operation of a university course, contingencies are such that students are never actually given the opportunity to train for and complete a specific goal which was felt to be highly valued. Such a development would be spotted by the "observation" part of the evaluation effort. This is not meant to imply that all courses at all times need involve only a mechanistic acting out of precisely preplanned activities. The development of new goals "spontaneously" during a course frequently occurs and should be evaluated just as any other.

Measuring Learning Outcomes. The purpose and means of course evaluation become clearer when we define teaching as facilitating student learning, or more broadly, causing changes in students toward highly valued goals. An instructor hopes to help a student move from where he is at the beginning of a course to where it is agreed he should be at the end. This

implies that we measure what he is able to do at the beginning and compare that to a measure of his knowledge and skills at the end of the course. The difference between the two is evidence of changes in the student. The arguments in favor of basing course and teacher evaluation on student learning effectiveness are presented more completely by Cohen and Brawer (1969). Suffice it to say that if universities are in the business of facilitating student learning, should student gains not be a most important criterion? As Anderson (1969) has said (in the quotation above) effectiveness is a crucial measure, although goals and other factors must be considered.

In the typical course evaluation system involving a questionnaire given at the end of a course, the problem of measuring effectiveness in terms of learning and changes in students has been largely ignored. It appears that we have been trying to measure effectiveness only indirectly all these years. Our questionnaires are aimed at describing what an instructor does or how his course is organized rather than the effects of the instructor or the course upon student learning. There is an implicit assumption that there is always a relationship between the dimensions rated on questionnaires and the outcomes of instruction.

To understand how far from an ideal system the questionnaire approach is, at least with respect to rating teachers, one

need only examine an experimental method such as that of Justiz (1968). He experimented in two different schools with teachers who were asked to teach two unfamiliar subject matter areas to randomly selected, unfamiliar students. The teachers were given one day to study the material themselves and to design a teaching strategy. Students were taught in groups of twenty, and were given tests over the material at the end of 30 minutes of instruction for each subject. Control groups of students were given the same tests without prior instruction. The difference between the average scores of experimental and control students provided evidence of learning in the experimental groups for each subject-matter area. Teachers were ranked according to the amount their students learned. These rankings were done separately for each of the two subject-matter areas. Thw two rankings were found to be significantly related, indicating that "teaching effectiveness" independent of subject matter or familiarity with students was reliably assessed.

Justiz' model for measuring effectiveness is quite specific and experimental. Its focus is on learning outcomes that could be assessed by short, objective tests. Certainly other models need to be developed for the university context. In any case, the major point being made here is that effectiveness can be measured more directly by investigating effects on

students rather than ratings by students.

How Do We Measure Learning? This is a frequent question that has emerged when the author has argued for basing course and instructor evaluation partially on student learning. Measurement of learning is perhaps too often narrowly defined as the use of objective tests, and many take exception to this. Clearly the methods of measuring learning are almost as varied as the kinds of students in a university. The design of appropriate learning measures is restricted only by the imagination of the instructor or evaluator. Measures based on projects, simulation excercises, or observations of students who are performing in a "natural setting" might be implied by the goals of a course.

Although the ideal situation would be when course evaluation included measures of learning which go beyond paper-and-pencil measures, even if the present exams and tests now used in college courses were seen as course-evaluation data an improvement might be made. Cronbach (1963) argued for the use of individual test questions as data for course improvement. Rather than using total test score, he suggested that individual questions be classified by type of content or objective and used as indicators of problems areas in instruction. The amount of time, effort and resources now going into the use of exams for grading purposes could and should be useful in

evaluation of the course as well as the student.

Cronbach (1963) has made some practical suggestions for evaluation that are particularly relevant to constraints such as large classes. He suggests the following:

1. Use "item sampling" techniques where there are many more criterion test questions than can be given to any one student. Sampling is the giving of a different test form to different samples of students.

2. For important objectives which can be measured only by complex or expensive criterion tests, e.g., tests of clinical history-taking by medical students, draw a random sample of students and observe them.

When such "evidence" of student learning has been compiled a report could be written discussing the findings and used for the variety of purposes course evaluation data has been used for in the past (e.g., given to Deans, Departmental Chairmen, published for students, given to colleagues who "inherit" the course in subsequent years, etc.).

If there is one most important area where new techniques and research are needed it is in the integration of learning outcomes in course evaluation.

A Word of Caution. In the use of any of the "tools" listed above for the evaluation of student learning and other outputs of the course, a problem of accuracy can exist. If these measures are only used at the end of the course it could be that students will show good performance or "changes" on them which may have been due to factors outside the course.

Such things as related courses, self-learning, and the communications media are possible outside sources which can facilitate changes even without the particular course. The chances of this may be low in courses which have unique content or train skills which would be difficult to learn elsewhere or without special equipment. Some further suggestions are:

1. In the case where students come <u>in</u> to the course with widely varying skills, measures can be taken at the beginning of the course. The pretest information is used to tell how much posttest performance is due to change during the course and how much to prior knowledge. This does not separate out all influences during the course, however.

2. Short segments of a course, e.g., one week, can be evaluated as a "mini-course". Outside factors would be less likely to occur in a shorter time period.

3. A more costly (in time and effort) approach would be to use a contrast group outside the course which could be given the same questionnaires or test (both pre- and post-measures). The group in the course should show greater changes than the contrast group. (Note: This approach is not completely free of problems when the two groups are systematically different. See Campbell and Stanley, 1963.)

The latter suggestion concerning the use of a contrast group, might suggest to the reader that traditional research designs are being considered an integral part of evaluation. A number of authors (e.g., Grobman, 1968; Saslow, 1970) have pointed out the differences between evaluation and research. The use of evaluation in service to decision making, such as the formative evaluation of a course (Scriven, 1967; Paulson, 1969), cannot be expected to contribute a great deal to the

science of learning. And, thus, the role of stringent control-group designs is in question. Cronbach (1963) argues against an emphasis on the comparative experiment in course improvement. Anderson (1969) and Scriven (1967) have discussed the value of such experiments, particularly when the objectives of two contrasted instructional packages are identical.

A System for Course Evalution

A number of aspects of an ideal course evaluation process have been discussed. Attention has been paid to the feasibility of actual changes in a course, finding information directly useful to decision makers, evaluating goals, and measuring learning outcomes. The integration of these activities within a system is depicted in Figure 1.

Insert Figure 1 about here

The system begins with a specific definition of the purpose of the course evaluation by all those who will use the information and judgements made public from it. Is the evaluation intended to help students, instructors or administrators (or all) to make decisions about the course? Will it be used as a measure of teaching ability? Will the emphasis be formative (the development of a new course) or summative (the testing of an established course)? Who is likely to examine the evaluation results? What will they do with it? Will their uses be valid? Control of the distribution of results from an evaluation is a crucial issue tied to the definition of purpose.

The time and money invested in course evaluation may be lost if the results it produces are not understood by or convincing to specific decision makers. Decision makers include

students who decide which courses to take, instructors who decide how to design them, colleagues of the instructor who decide whether the course should be a prerequisite for theirs, administrators and planners who decide on curriculum and allocation of resources. Perhaps the ultimate decision makers are the government representatives or the people themselves who contribute to the financial support of universities. As the emphasis on accountability in education increases, more and more attention will need to be paid to the latter. It is critically important that an initial attempt is made to find out what types of information or judgements these people find useful and convincing. This will be learned over time by actual try-out, but a start can be made early.

An assessment of whether real and significant changes will be possible if evaluation is conducted must be guessed. If the course depends on unchangeable resources, people or traditions, perhaps it is best to stop here before anymore time is wasted. Energy should go immediately into building the possibility of change into a department or university. This is, of course, a statement of the ideal and it may be that the evaluation would contribute to change although not bring it about directly.

Next, the various subgroups served by the course are identified. Many of these will be the decision makers already

found. The purpose of this is to find out who should be polled about the course. These groups of students, colleagues, administrators and others will be good sources for statements of what is expected of and accomplished in the course.

The intended goals of the course are now specified. The instructor lists what he hopes to accomplish in the course before it starts. Students indicate what they expect and how the course fits into their career and personal goals as they understand it. These intended goals may in fact not be the actual goals that develop as the course is run. However, evaluation of them is a type of formative evaluation which helps the initial design of the course. They are evaluated in terms of clarity and meaningfulness as was discussed in an earlier section of this paper. If they do not prove acceptable, they can be changed immediately or the students and decision makers need to be polled again for purposes of clarification. One could imagine an "iterative interview" technique (Saslow, 1971) in which ,first ,representative people are asked to explain the purposes and goals of a course. Then a tentative list is compiled. This list is returned to the interviewees for further clarification, and polling.

Next an initial list of intended inputs, processes, and outputs for a course is drawn up. This information will guide development of appropriate observation techniques. What students

is the course aimed at? How many will there be? What classrooms, field locations or environments will the students be in contact with? What is the proposed set of teaching methods to be used? Are media involved? What kinds of new skills or knowledge does the course intend to facilitate? All of these questions imply objects or activities that need to be observed and evaluated.

Observation methods, test, checklists, simple frequency counts and other measures which can be pre-designed are now constructed.

With a purpose for the evaluation, a description of the students served by the course, a list of intended goals, and a design for observation, it should now be possible to estimate the approximate cost of the evaluation. If the cost/effectiveness of the evaluation in terms of actual change that it may bring is unacceptable, again a consideration to stop here should be made. This is a difficult decision to make because as Paulson (1969) has said, "The costs of evaluation are much easier to determine than the costs of ignorance of such information." The decision to not evaluate and to continue a course in an imperfect form may have certain costs implied, e.g., wasted resources or student time. It may be, however, that certain parts of the evaluation are cost/effective. For example, observations made during the course may help

change the course immediately, but a study of graduates from five years previously may be more costly with an unknown payoff in course improvement.

Observations and measures are now taken on the course as it begins its operation. Emphasis is now shifted to actual goals which might be reflected only in examinations or the directions taken by discussion groups. Actual inputs may be different from intended as minor crises come up requiring different resources, or, in terms of student input, as some students drop out. Actual processes are almost certain to be different from those intended as teaching methods are applied to individual students for the first time. As actual goals, inputs or processes change, the output expected of or actually accomplished by students may change. Of course, measurement devices and observation schemes will suffer under the pressure of frequent changes during a course. The closer the original plan for the course can match what actually happens the more time the evaluator will have to prepare and the greater the chances are for appropriate measures being used.

It is most important at this point in the system that the question be asked, "Were the observation techniques just used appropriate and valid measures of what we were trying to measure?" Trying out exams, simulation excercises or unobtrusive measures (Webb, et. al., 1966) for the first time on individual students

may reveal unanticipated errors that need changing.

Once observations on at least a portion of the course are complete the "judgement" part of evaluation can take place. Actual accomplishments of students are compared with those intended or those reflected in actual goals. For example, a lab course in chemistry may have intended for students to master the execution of small experiments. Due to a cut in budget, lab equipment was incomplete so the goal had to be changed to helping students master only the design of small experiments. Papers describing experimental designs have been collected and assessed. Judgements are made as to what extent students mastered the elements of good design.

Findings, data, and evaluative judgements are written up in report form, published or otherwise distributed to those people who make decisions about improving the course. Attention must be paid here to the form of information presentation that decision makers really understand. Statistical analyses may or may not be interpretable or convincing to these people. The author has personally seen many computer print-outs of results discarded as uninterpretable by instructors. The reader is refered to Paulson (1970) for further comments and guidelines on this matter.

A mini-experiment can be run now by taking evaluation reports to decision makers and discussing results with them.

An attempt is made to see whether the information is understood and convincing. Also, the evaluator might have to sit back temporarily and wait to see visible signs of action (or inaction) taken on the basis of his findings. If inaction occurs perhaps the report needs to be revised, data presented differently or summarized more accurately. Perhaps a more serious redesign of observation techniques is implied.

Once evaluation data is found useful, strategies for actually changing the course (if needed) can be designed. This may involve planning meetings, further training of staff, search for new instructional materials or changes in administrative rules such as those governing the form of exams or grading practices.

Course improvement may involve changing some of the goals, aiming the course at a more well defined subgroup of students, changing teaching methods, or changing the standards of what students are expected to learn and accomplish. This process can involve the instructor of the course in the growing area of instructional design (e.g., Briggs, 1970).

Implementation Models

It will be of interest to examine some examples of the implementation of the kind of course evaluation presented in this paper. Several models of partial or complete systems, some being tried now and others proposed, are described below.

Use of Course Exams. One model which implements only the measuring of learning outcomes as part of course evaluation is depicted in a case study in which the author served as an evaluation consultant to a course. The author tried to move the course instructor towards integrating regular tests into evaluation. Also, an attempt was made to design questionnaire measures which would help the instructor make actual changes in the materials and grading procedures of the course. The course was experimental, using a method called modular instruction. Modular instruction involves the separation of a course into major concepts, topics or task units. Self-instructional packages are developed for each unit and called "modules". Each of the 14 modules in the course contained a study guide with objectives and reading, a taped commentary and slides. The students worked through the modules at their own pace and met with a teaching assistant at a "drop-in center" whenever necessary. They presented themselves for testing (group oral exam and written quizes) whenever they completed a module and felt ready. Testing could be re-taken a limited number

of times (different test forms used). The case study of this course is included below, and is written in the first person from the viewpoint of the evaluation consultant.

The instructor initiated contact with me for purposes of examining the questionnaire he hoped to use to evaluate his course. His contact was certainly encouraged by a strong need to make an evaluation report for the agency which funded his new course.

I visited his assistant working at the drop-in centre to discuss their questionnaire and to find out more about the course. They gave completed questionnaires to me to read in hopes that I might be able to find out something about the course that they could not conclude themselves by reading them. The information level of this questionnaire was very low, and did not lead to noticeable changes in the course.

I arranged the next meeting with the instructor himself in hopes of moving him more in the direction of formative and summative evaluation based on module tests in addition to questionnaire data. We had a general discussion about purposes of his evaluation and I made some suggestions along these lines that eventually met with his approval. Then, we worked out a plan in outline form that looked something like:

1. Give short questionnaire now for administrative decisions and decisions about audio tapes.

2. Design evaluation plan for module tests and get the assistant to tabulate test results in proper form.

3. Design final course questionnaire for the end of the course.

We started with the first and drafted the questionnaire right at that meeting. We tailored questions specifically to things he wanted information on to change. This second questionnaire is attached. His areas of possible change are (a) number of modules required for a grade, (b) reductions of work load by (a) and by reducing supplementary readings, and (c) improvement of audio tapes still to be made for the remaining modules.

In a second meeting with the course assistant, some tangible changes made on the basis of the questionnaire were discovered.

The number of modules required was reduced to 8 instead of 10 to cut down work load. Supplementary reading was reduced. Audio tape music and pauses were changed on new tapes. So, it seems that the questionnaire really served a feedback function.

The course assistant then worked on what are to my mind most important tasks:

1. Computing the average final posttest score for each module. Students were required to get 80% (8 out of 10 questions based on objectives) to pass a module. if the average is higher than 80%, we have evidence of "more-than-expected" in student learning.

2. Identifying students who have taken more than one posttest for each module (they take these quizes until they pass). She computes the average of all first-test scores and compares that to the average of final test scores. The difference is evidence of student learning.

3. Identifying students who passed each module with only one posttest and computing the proportion of such people for each module. This proportion has apparently been increasing and is evidence that students are adapting to the contingencies in the system - perhaps even working more efficiently as the course has progressed. Initial guess is that this proportion has increased from 30% to 80% from first to current modules.

These evaluation procedures can be useful for both summative and formative evaluation. Some of the data will show differences between modules indicating where some should be made less or more difficult - e.g., proportion of students passing a module on the first posttest varies within the overall increasing trend.

Further formative evaluation procedures which are planned for the future are based on item analyses of posttests to identify specific parts of modules for revision. Since students were given one of five possible posttest quizes randomly for each module the number of students will be as small as five for some item analyses. However, the data will be at least suggestive eventhough not definitive.

Golden West Model. A model of continuous evaluation was described in an article by Cohen and Brawer (1969) as being implemented at Golden West College in California. The basic structure of the model is to have regular "help sessions" during the year for the instructor. The instructor, a department chairman or supervisor, and an instructional specialist meet to discuss course objectives and data on student learning. Instructors who are unfamiliar with specifying objectives are given training.

A revision of this plan with more detailed steps might be as follows:

Step 1: Training in specifying objectives.

Step 2: Evaluation of objectives during initial help sessions.

Step 3: Training in measurement of objectives.

Step 4: Collection of data on student progress toward objectives.

Step 5: Evaluation and interpretation of data during regular help sessions.

Step 6: Feedback and course improvement. Resources are allocated to help the instructor to cause student learning.

Course Improvement Team. Another model involves the training and dissemination of evaluation expertise into university departments. Rather than have an evaluation center which disseminates only questionnaires, and tabulated results, a center could train "satelite" evaluators in university departments. The combination of subject matter expert and evaluation consultant is an interesting new role which might be developed. The complexity of a thorough course evaluation system requires availability of this expertise. An analysis of the roles and tasks involved in designing and operating a course shows the need for such people. The evaluation and improvement of instruction can be seen as a continuous research effort in which "instructors" play a number of different roles such as:

1. Objectives designer. One person has primary responsibility for designing good, measurable objectives. This can be a full job if it involves (a) looking for reference material on objectives, (b) trying out objectives on samples of students, (c) surveying graduates of previous years to see what objectives are most appropriate, (d) evaluating current objectives.

2. Evaluation designer. May be the same person who designs objectives, but not necessarily, particularly in beginning stages of course design. He develops test forms, tries them out on student samples, and revises them. He

constantly asks the gnawing question "are we <u>really</u> measuring what we want the students to do?"

3. <u>Process managers</u>. These are instructors who actually appear in class, give tests and talk with the students enrolled in the course. Some may be "lecture specialists", others may be "discussion specialists". Their functions may overlap with the following people.

4. <u>Materials designers</u>. These people find and/or design instructional material for the course. They analyze the steps involved in students achieving the objectives and construct learning materials for each. They tryout and revise materials on sample students prior to use. This team could meet at regular intervals to share findings, discuss data collection methods, and interpret results. Each writes a report from his viewpoint on student learning in the course. Each coauthors an overall evaluation report for the course.

<u>Student Change Team</u>. Another model might involve a type of educational program that does not resemble current courses. This is a research team approach which focuses on change in the individual student. The approach might be adapted from programs such as that carried out by Fox (1962) for the training of students in study habits, and is basically a clinical, behavior modification model. The model is based on the idea of finding for each student the appropriate behavior to be

changed (what does the student and the instructor feel he should be able to do by the end of the course?) Then data is kept on his progress on a continuing basis. The student helps the data collection process by keeping regular records of his studying. A central office with non-professional staff could be maintained as a daily "check-in-center". The "course" is transformed into a resource course where instructional materials are available and a "consultant team" is available to study each student's own progress and make recommendations to him.

The following would be aspects of such a program:

1. Initial interviews which specify the instructional problem and the learning change contracted with each student.

2. Collection of data on the current performance of the student (called "baseline") to be used for future comparison.

3. Training involving learning materials and feedback.

4. Analysis and adjustment of rewards and reinforcements that help or hinder learning progress. The student may elect self-management training (e.g., Murdock, 1971).

Roles on the team for this model may include an initial interviewer and contact, a data recorder-analyst, a contingency manager and materials and training director. The advantage of such a model is that evaluation of the student and the program are an inseparable part of the program. Each student is a case study, reviewed regularly. It assumes that the

business of education is to increase the frequency of certain behaviors in students. These may be such behaviors as, "Talking about the subject of _____", or "Being able to build a model of _____", or "Reading lots of _____", or "Doing those things a professional _____ does". In order to increase, create or shape these new behaviors evaluation of progress is essential. When progress is slow the program must be examined.

This behavior-change model is the most radical discussed so far in that it implies a dramatic change in the organization and staffing of a university course. For this reason it is important to consider the following.

A Final Note on Incentives

No complex skill or behavior will be maintained in an organization if it is not highly prized by the organization or the people in it. This paper can suggest endless schemes to redesign evaluation programs in universities, all to no avail if priorities are not changed or incentives provided for instructors to actually use them. At this time in history there is little evidence that universities reward the kind of thorough evaluation of their courses advocated in this paper. Priorities do not currently allow a professor to devote the time required to adequately evaluate his courses

or train himself to be an evaluation expert.

Evaluation in general appears to be something to which much lip service is paid but few incentives are provided for. In fact, evaluation can be absolutely punishing to the evaluator or instructor in a course. He may find that a course is in dismal shape and that students are accomplishing little. These experiences tend to drive people away from evaluation.

So, if there is one other element which could be added to a system of course evaluation (and there must be others) it would be some incentives for the evaluator or course designer. In fact, the author in reviewing the task of writing this paper has come to the brink of the feeling that the important task is not the presentation of new systems or models but the changing of the structure and priorities of the modern university towards accountability.

References

Anderson, R.C. The comparative field experiment: An illustration from high school biology. Proceedings of the 1968 Invitational Conference on Testing Problems, Educational Testing Service, Princeton, New Jersey, 1969.

Briggs, L.J. Handbook of procedures for the design of instruction. Pittsburgh: American Institutes for Research, 1970.

Campbell, D.T. and Stanley, J.C. Experimental and quasi-experimental designs for research on teaching. In: N.L. Gage (Ed.) Handbook of Research on Teaching, Chicago: Rand-McNally, 1968.

Cohen, A.M. and Brawer, F.B. Measuring faculty performance. American Association of Junior Colleges, 1969.

Cohen, S.H. and Berger, W.G. Dimensions of student ratings of college instructors underlying subsequent achievement on course examinations. Proceedings of the 78th Annual APA Convention, 1970, pp. 605-606.

Cronbach, L.J. Evaluation for course improvement. Teachers College Record, 1963, 64, 672-683.

Fox, L. Effecting the use of efficient study habits. Journal of Mathetics , 1962, 75-86.

Geis, G.L. Premature instruction. Educational Technology, April, 1970, Vol. 10, No. 4, 24-30.

Grobman, H. What to evaluate. Chapter 2 in: Evaluation Activities of Curriculum Projects, AERA Monograph Series on Curriculum Evaluation, Volume 2, Chicago: Rand-McNally, 1968.

Justiz, T.D. A method for identifying the effective teacher. Doctoral dissertation, University of California, Los Angeles, 1968. (University Microfilms Order No. 69-5321).

Murdock, E.E. Self-Management: A guide to more effective study. San Rafael, California: Independent Learning Systems, 1971.

Nelson, F. and Paulson, C.F. Behavioral objectives. In: Crawford, J. (Ed.), National Research, Evaluation and Development (C.O.R.D.) Manual, 2nd Edition, Teaching Research, Monmouth, Oregon, 97361, 1969.

Paulson, C.F. Evaluation of instructional systems. In: Crawford, J. (Ed.) National Research Evaluation and Development Training (C.O.R.D.) Manual, 2nd Edition, Teaching Research Division, Oregon State System of Higher Education, Monmouth, Oregon, 97361, 1969.

Paulson, C.F. (Ed.) A strategy for evaluation design. Teaching Research Division, Oregon State System of Higher Education, Monmouth, Oregon, 97361, 1970.

Popham, W.J. and Baker, E.I. Establishing instructional goals. Englewood Cliffs, New Jersey: Prentice-Hall, 1970.

Saslow, M.G. Establishing the purpose for evaluation. In: Paulson, C.F. (Ed.) A strategy for evaluation design. Teaching Research Division, Oregon State System of Higher Education, Monmouth, Oregon, 97361, 1970.

Saslow, M.G. Newgate: Problem solving approach for agency assistance. Mimeo, Teaching Research Division, Oregon State System of Higher Education, Monmouth, Oregon, 97361, 1971.

Scriven, M. The methodology of evaluation. In: Tyler, R.W., Gagné, R.M., and Scriven, M. (Eds.), Perspectives of curriculum evaluation, AERA Monograph Series on Curriculum Evaluation, Volume 1, Chicago: Rand-McNally, 1967.

Stake, R.E. The countenance of educational evaluation. Teachers College Record, 1967, 68, 523-540.

Stufflebeam, D.L. Toward a science of educational evaluation. Educational Technology, 1968, 8, 5-12.

Webb, E.J., Campbell, D.T., Schwartz, R.D., and Sechrest, L. Unobtrusive measures: Nonreactive research in the social sciences. Chicago, Illinois: Rand-McNally, 1966.

Yelon, S.L. and Scott, R.O. A strategy for writing objectives. Dubuque: Kendall-Hunt, 1970.

FIGURE 1

A System for Course Evaluation